Modularity clustering is force-directed layout

Andreas Noack
Institute of Computer Science, Brandenburg University of Technology, 03013 Cottbus, Germany
(Dated: December 9, 2008)

Two natural and widely used representations for the community structure of networks are clusterings, which partition the vertex set into disjoint subsets, and layouts, which assign the vertices to positions in a metric space. This paper unifies prominent characterizations of layout quality and clustering quality, by showing that energy models of pairwise attraction and repulsion subsume Newman and Girvan's modularity measure. Layouts with optimal energy are relaxations of, and are thus consistent with, clusterings with optimal modularity, which is of practical relevance because both representations are complementary and often used together.

PACS numbers: 89.75.Hc, 02.10.Ox

I. INTRODUCTION

Many systems of scientific or practical interest are decomposable into subsystems with strong internal and relatively weak external interactions [1]; for example, there are groups of friends or collaborators in social networks, sets of topically related documents in hypertexts, or blocs of interlocked countries in international trade. If systems are modeled as networks, with the system elements as vertices and their interactions as edges, then each subsystem corresponds to a so-called *community*, a set of vertices with dense internal connections but sparse connections to the remaining network.

Two widely used representations of networks are *layouts*, which assign the vertices to positions in a metric space, and *clusterings*, which partition the vertex set into disjoint subsets. Both representations can group densely connected vertices, by placing them at nearby positions or in the same cluster, and separate sparsely connected vertices, by placing them at distant positions or in different clusters, and can thus naturally reflect the community structure. Requirements like the grouping of densely connected vertices are often formalized as mathematical functions called *quality measures*, and the optimization of quality measures is a common strategy for the computation of both layouts [2, 3] and clusterings [4, 5, 6, 7]. Despite these commonalities, and although layouts and clusterings are often used together as complementary representations of the same network, there is no coherent understanding of layout quality and clustering quality.

This paper unifies Newman and Girvan's modularity [8], a popular quality measure for clusterings, with energy models of pairwise attraction and repulsion between vertices (e.g., [2, 3]), a widely used class of quality measures for layouts. After an introduction of the quality measures in Sec. II, Sec. III shows that layouts with optimal energy and clusterings with optimal modularity represent the community structure similarly, and Sec. IV demonstrates that modularity actually *is* an energy model of pairwise attraction and repulsion, if clusterings are considered as restricted layouts. Section V discusses the application of these results for computing consistent clusterings and layouts.

II. ENERGY MODELS AND MODULARITY

Quality measures for representations of networks formalize what is considered as a *good* representation, and allow to compute good representations automatically using optimization algorithms. Mathematically, a quality measure maps network representations to real numbers, such that larger (or smaller) numbers are assigned to better representations, and the best representations correspond to maxima (or minima) of the measure. This section introduces two widely used quality measures, namely energy models based on pairwise attraction and repulsion for layouts, and Newman and Girvan's modularity measure for clusterings.

To obtain uniform and general formulations, both measures are defined for *weighted* networks. In a weighted network, each vertex v has a nonnegative real *vertex weight* w_v, and each unordered vertex pair $\{u, v\}$ (including $u = v$) has a nonnegative real *edge weight* $w_{\{u,v\}}$. Intuitively, a vertex (or edge) of weight k can be thought of as a chunk of k vertices (or edges) of weight 1. The commonly studied *un*weighted networks correspond to the special case where the edge weights are either 0 (no edge) or 1, and the vertex weights are 1.

A. The (a, r)-energy model for layouts

A *d-dimensional layout* p of a network maps each vertex v to a position p_v in $\mathbb{R}^d$; it thereby assigns a *distance* to each vertex pair $\{u, v\}$, namely the Euclidean distance $\|p_u - p_v\|$ between the respective vertex positions. So-called *energy models* are an important class of quality measures for layouts. In general, *smaller* energy indicates better layouts. Because force is the negative gradient of energy, energy models can also be represented as force systems, and energy minima correspond to force equilibria. For introductions to energy-based or force-directed layout, see Refs. [2, 3].

The most popular energy models for general undirected networks are either similar to stress functions of multidimensional scaling [9], or represent force systems of pairwise attraction and repulsion between vertices. Mod-

els of the former type (e.g., [10]) enforce that the distance of each vertex pair in the layout approximates some prespecified distance, most commonly the length of the shortest edge path between the vertices. They will not be further discussed, because their layouts reflect these path lengths rather than the community structure.

In models of the latter type, adjacent vertices attract, which tends to group densely connected vertices, and all pairs of vertices repulse, which tends to separate sparsely connected vertices. The strengths of the forces are often chosen to be proportional to some power of the distance. Formally, for a layout p and two vertices u, v with $u \neq v$, the attractive force exerted on u by v is

$$w_{\{u,v\}} \, \|p_u - p_v\|^a \, \overrightarrow{p_u p_v} \ ,$$

and the repulsive force exerted on u by v is

$$w_u w_v \, \|p_u - p_v\|^r \, \overrightarrow{p_v p_u} \ ,$$

where $\|p_u - p_v\|$ is the distance between u and v, $\overrightarrow{p_u p_v}$ is the unit-length vector pointing from u to v, and a and r are real constants with $a > r$.

The condition $a > r$ ensures that the attractive force between connected vertices grows faster than the repulsive force, and thus prevents infinite distances except between unconnected components. For most practical force models holds $a \geq 0$ and $r \leq 0$, i.e., the attractive force is non-decreasing and the repulsive force is non-increasing with growing distance. In the widely used force model of Fruchterman and Reingold [11], $a = 2$ and $r = -1$.

By exploiting that force is the negative gradient of energy, the force model can be transformed into an energy model, such that force equilibria correspond to (local) energy minima. For a layout p and constants $a, r \in \mathbb{R}$ with $a > r$, the (a, r)-*energy* is

$$\sum_{\{u,v\}\colon u \neq v} \left(w_{\{u,v\}} \frac{\|p_u - p_v\|^{a+1}}{a+1} \ - \ w_u w_v \frac{\|p_u - p_v\|^{r+1}}{r+1} \right), \tag{1}$$

where $\frac{\|p_u - p_v\|^{-1+1}}{-1+1}$ must be read as $\ln \|p_u - p_v\|$ (because x^{-1} is the derivative of $\ln x$). The $(1, -3)$-energy model has been proposed by Davidson and Harel [12], and the $(0, -1)$-energy model is known as LinLog model [13, 14].

B. The modularity measure for clusterings

A *clustering* p of a network partitions the vertex set into disjoint subsets called *clusters*, and thereby maps each vertex v to a cluster p_v. Proposals of quality measures for clusterings are numerous and scattered over the literature of diverse research fields; surveys, though non-exhaustive, are provided by Refs. [5, 6, 14, 15].

One of the most widely used quality measures was introduced by Newman and Girvan, and is called modularity. It was originally defined for the special case where the edge weights are either 0 or 1 and the weight of each vertex is its degree [8], and was later extended to networks with arbitrary edge weights [16]. (The *degree* of a vertex is the total weight of its incident edges, with the edge weight from the vertex to itself counted twice.) Generalized to arbitrary vertex weights, the *modularity* of a clustering p is

$$\sum_{c \in p(V)} \left(\frac{w_{\{c,c\}}}{w_{\{V,V\}}} - \frac{\frac{1}{2} w_c^2}{\frac{1}{2} w_V^2} \right), \tag{2}$$

where V is the set of all vertices in the network, and $p(V)$ is the set of clusters; the weight functions are naturally extended to sets of vertices or edges: $w_{\{c,c\}}$ is the total edge weight within the cluster c, and w_c is the total weight of the vertices in c.

Intuitively, the first term of the modularity measure is the *actual* fraction of intra-cluster edge weight. In itself, it is not a good measure of clustering quality, because it takes the maximum value for the trivial clustering where one cluster contains all vertices. This is corrected by subtracting a second term, which specifies the *expected* fraction of intra-cluster edge weight in a network with uniform density. Thus modularity takes positive values for clusterings where the total edge weight within clusters is larger than would be expected if the network had no community structure.

C. Optimization algorithms

Finding a minimum-energy layout or a maximum-modularity clustering of a given network is computationally hard; in particular, modularity maximization was recently shown to be NP-complete [17]. In practice, energy and modularity are almost exclusively optimized with heuristic algorithms that do not guarantee to find optimal or near-optimal solutions.

An extensive experimental comparison of energy minimization algorithms for network layout was performed by Hachul and Jünger [18]; however, most of the examined algorithms make fairly restrictive assumptions about the optimized energy model. More general and reasonably efficient is the force calculation algorithm by Barnes and Hut [19], whose runtime is $O(m + n \log n)$ per iteration for a network with m edges (with nonzero weight) and n vertices (assuming that the number of dimensions is small and the vertex distances are not extremely nonuniform). The number of iterations required for convergence typically grows sublinearly with n.

Clustering algorithms for networks are surveyed in Refs. [4, 6, 7, 20]. A relatively fast yet very effective heuristic for modularity maximization is agglomeration by iteratively merging clusters (starting from singletons), combined with single-level [21] or multi-level [22] refinement by iteratively moving vertices; an efficient implementation requires a runtime of $O(m \log^2 n)$ (assuming $O(\log n)$ hierarchy levels in agglomeration and $O(\log n)$ iterations through all vertices per level in refinement).

III. ENERGY MODELS AND MODULARITY REVEAL COMMUNITIES

A set of vertices is called a *community* if the density within the set is significantly larger than the density between the set and the remaining network. The *density between* two disjoint sets of vertices T and U is intuitively the quotient of the actual edge weight and the potential edge weight between T and U; formally, it is defined as $\frac{w_{\{T,U\}}}{w_T w_U}$, where w_U is the total weight of the vertices in U, and $w_{\{T,U\}}$ is the total edge weight between T and U. Similarly, the *density within* a vertex set U is $\frac{w_{\{U,U\}}}{w_U^2/2}$. (This generalizes standard definitions of density from graph theory [23] to weighted networks with self-edges.)

Existing theoretical results, which will be summarized and extended in this section, already show that the community structure of a network is reflected in layouts with optimal (a,r)-energy (for certain values of a and r) and in clusterings with optimal modularity. What has previously escaped notice is the striking analogy: *The separation of communities in an optimal layout is inversely proportional to (some power of) the density between them, and the separation of communities in an optimal clustering reflects whether the density between them is smaller than a certain threshold.* As an important limitation, the result for layouts will be derived only for two communities, and cannot be expected to hold precisely for more communities. Therefore, the consistency of (a,r)-energy layouts and modularity clusterings will be revisited in Sec. V, after further evidence has been presented in Sec. IV.

In what appears to be the only previous work that formally relates energy-based layout to modularity clustering [14], we did not established similarities between optimal layouts and optimal clusterings, but only noted that the modularity measure is mathematically similar to the density (called normalized cut in [14]), as both normalize the actual edge weight with a potential or expected edge weight.

A. Representation of community structure in layouts with optimal (a,r)-energy

This subsection discusses how the distances in a layout with optimal (a,r)-energy can be interpreted in terms of the community structure of the network, and how this interpretation depends on the parameters a and r.

For the simple case of a network with two vertices, the minimum-energy layouts can be computed analytically (Theorem 3 in [14]). If the vertices u and v have the distance d, the (a,r)-energy is

$$U(d) \;:=\; w_{\{u,v\}}\frac{d^{a+1}}{a+1} \;-\; w_u w_v \frac{d^{r+1}}{r+1}\;.$$

The derivative of this function is 0 at its minimum d_0, thus

$$\begin{aligned} 0 \;=\; U'(d_0) \;&=\; w_{\{u,v\}} d_0^a \;-\; w_u w_v d_0^r \\ d_0 \;&=\; \left(\frac{w_{\{u,v\}}}{w_u w_v}\right)^{-\frac{1}{a-r}}\;. \end{aligned} \tag{3}$$

Thus the distance of the two vertices in a layout with optimal (a,r)-energy is the $-\frac{1}{a-r}$th power of the density between the vertices. In particular, the distance is the inverse density if $a-r=1$, and the distance is almost independent of the density if $a-r \gg 1$. This impact of $a-r$ on the representation of the community structure is illustrated for a larger network in Fig. 1.

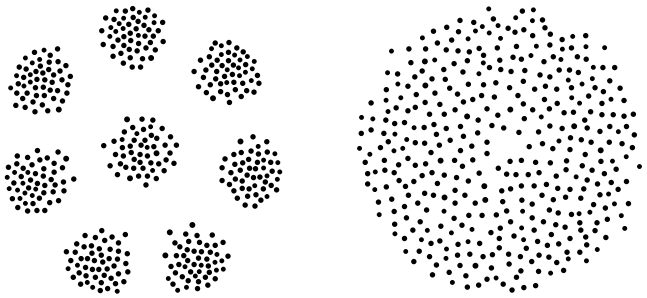

FIG. 1: Layouts with small LinLog energy ($a-r=1$) and with small Fruchterman-Reingold energy ($a-r=3$) of a pseudo-random network with eight clusters (intra-cluster density 1.0, expected inter-cluster density 0.2).

Replacing the edge $\{u,v\}$ with two edges $\{u,t\}$ and $\{t,v\}$, where t is a new vertex with weight 0, increases the optimal distance between u and v by a factor of $2^{a/(a-r)}$. Because the (a,r)-energy is only defined for $a-r>0$, the factor is 1 if $a=0$, and greater than 1 if $a>0$. This result has a significant implication, given that the addition of t increases the path length between u and v (from 1 to 2 edges) without changing the density: The optimal distance of u and v depends only on the density, and not on the path length, if $a=0$ (as in the LinLog energy model), and increases with the path length if $a>0$.

The results for networks with two or three vertices can be generalized, at least as approximations, to larger networks. In a network with clear communities, for example, the density within the communities is (by definition) much greater than the density between the communities, and thus the intra-community distances in an optimal layout are much smaller than the inter-community distances (unless $a-r$ is very large). This can be approximated by assuming that the vertices of each community have the same position, and thus by considering each community as one big vertex. For networks with more than two communities, Eq. (3) cannot be expected to hold precisely for all pairs of communities, because this would often imply distances that violate the triangle inequality. Nevertheless, the qualitative reasoning generalizes: Distances are less dependent on densities for large $a-r$, and less dependent on path lengths for small a.

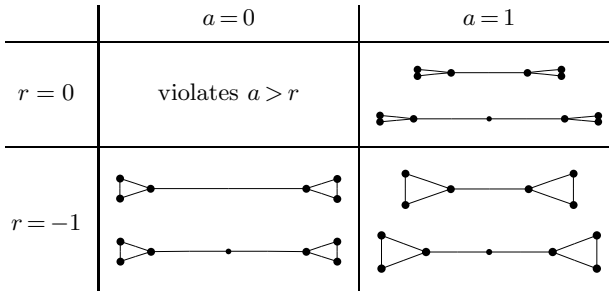

FIG. 2: Layouts with optimal (a,r)-energy for different values of a and r. All vertices and edges have weight 1, except for the small vertex between the triangles which has weight 0.

Figure 2 illustrates the impact of the parameters a and r for two simple networks: For $a-r>1$ (bottom right), the two triangles are less clearly separated than for $a-r=1$ (bottom left and top right), and only for $a=0$ (left), the path length between the triangles does not affect their distance.

Figure 3 summarizes the results of this subsection.

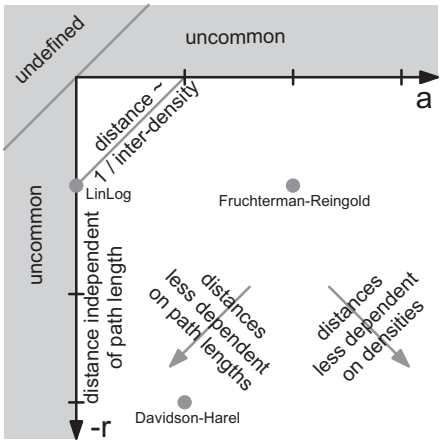

FIG. 3: Impact of the parameters a and r on the optimal layouts of the (a,r)-energy model.

B. Representation of community structure in clusterings with optimal modularity

Reichardt and Bornholdt [24] observed that in a clustering with maximum modularity, the density between any two clusters is at most the density within the entire network, and the density between any two subclusters obtained by splitting a cluster is at least the density within the network. (Clusters may still have a smaller density than the network, essentially because vertices without self-edges decrease the density within their cluster but cannot be split.) The argument is simple: Joining two clusters c and d with $c\neq d$ increases the modularity by

$$\frac{w_{\{c,d\}}}{w_{\{V,V\}}} - \frac{w_c w_d}{\frac{1}{2}w_V^2} ,$$

which is positive if and only if

$$\frac{w_{\{c,d\}}}{w_c w_d} > \frac{w_{\{V,V\}}}{\frac{1}{2}w_V^2} ,$$

i.e., if the density between c and d is greater than the density within the network. In a clustering with maximum modularity, neither joining nor splitting clusters may increase the modularity, which yields the claim.

These observations imply that the granularity of clusterings with maximum modularity depends on the overall density within the network, which may be undesirable for some applications. For example, if the density within the network is sufficiently small, then two dense subnetworks connected by only one light-weight edge are joined into a single cluster, instead of forming two separate clusters [25]. Similarly, doubling a network (by adding a second copy of the same network) halves its density, and thus generally coarsens the optimal clustering instead of preserving it [17]. Because such granularity-related issues are specific to discrete representations like clusterings, they provide a major motivation for the supplementary (and sometimes even exclusive) use of continuous representations like layouts.

IV. ENERGY SUBSUMES MODULARITY

Modularity can be considered as a special case of (a,r)*-energy.* The first subsection formally derives this result, and the second subsection explains how this derivation is facilitated by the definitions of (a,r)-energy and modularity in Sec. II, which generalize previous definitions from the literature.

A. Transformation of modularity into (a,r)-energy

The modularity of a clustering p was defined in Sec. II B as

$$\sum_{c\in p(V)} \left(\frac{w_{\{c,c\}}}{w_{\{V,V\}}} - \frac{\frac{1}{2}w_c^2}{\frac{1}{2}w_V^2} \right) ,$$

i.e., as the difference of the actual fraction of intra-cluster edge weight and the expected fraction of intra-cluster edge weight.

Because each edge is either intra-cluster or inter-cluster, the fraction of intra-cluster edge weight and the fraction of inter-cluster edge weight add up to 1:

$$\sum_{c\in p(V)} \frac{w_{\{c,c\}}}{w_{\{V,V\}}} + \sum_{\{c,d\}\subseteq p(V):\, c\neq d} \frac{w_{\{c,d\}}}{w_{\{V,V\}}} = 1 ;$$

similarly, the corresponding expected fractions add up to 1. Thus the modularity of p can be written in terms

of inter-cluster edge weights as

$$\sum_{\{c,d\}\subseteq p(V)\colon c\neq d} \left(-\frac{w_{\{c,d\}}}{w_{\{V,V\}}} + \frac{w_c w_d}{\frac{1}{2}w_V^2} \right) = -\sum_{\{u,v\}\subseteq V\colon p_u\neq p_v} \left(\frac{w_{\{u,v\}}}{w_{\{V,V\}}} - \frac{w_u w_v}{\frac{1}{2}w_V^2} \right).$$

Let k be the number of clusters in p. Without changing the modularity of p, the k clusters can be considered as positions in $\mathbb{R}^{k-1}$, such that each pair of different clusters has the distance 1. (Intuitively, the k clusters form the corners of a regular $(k-1)$-simplex with edge length 1; a $(k-1)$-simplex is the $(k-1)$-dimensional analogue of a triangle.) Then the clustering p is a $(k-1)$-dimensional layout, and the modularity of p can be rewritten as

$$-\sum_{\{u,v\}\subseteq V\colon p_u\neq p_v} \left(\frac{w_{\{u,v\}}}{w_{\{V,V\}}} \|p_u - p_v\| - \frac{w_u w_v}{\frac{1}{2}w_V^2} \|p_u - p_v\| \right).$$

The condition $p_u \neq p_v$ of the sum can be dropped or replaced with $u \neq v$, because it excludes only vertex pairs $\{u, v\}$ with $\|p_u - p_v\| = 0$.

Because the distances between the vertices are 0 or 1, the modularity of p equals

$$-\sum_{\{u,v\}\colon u\neq v} \left(\frac{w_{\{u,v\}}}{w_{\{V,V\}}} \|p_u - p_v\|^{a+1} - \frac{w_u w_v}{\frac{1}{2}w_V^2} \|p_u - p_v\|^{r+1} \right)$$

for all $a, r \in \mathbb{R}$ with $a > -1$ and $r > -1$. This is the negative (a, r)-energy, except for the constant factors in the attraction term and the repulsion term, which change only the scaling of the optimal layouts.

B. Prerequisites of the transformation

The transformation of modularity into (a, r)-energy in the previous subsection is based on the definitions of the measures in Sec. II, which generalize previous definitions from the literature in several respects.

First, the goal of most energy-based layout techniques is to produce easily readable box-and-line visualizations, which differs from and even conflicts with producing faithful representations of the community structure. The classic energy models of Eades [26], Fruchterman and Reingold [11], and Davidson and Harel [12] primarily reward the conformance to aesthetic criteria like small edge lengths and uniformly distributed vertices, and thus often prevent the clear separation of sparsely connected vertices and the clear grouping of densely connected vertices (see Fig. 1). The design and evaluation of energy models with the explicit purpose of representing the community structure started only recently with the LinLog model [13, 14]. Technically, the classic energy models are, or are similar to, instances of the (a, r)-energy model where the difference $a - r$ is fixed and too large; the (a, r)-energy model is parameterized with this difference.

Second, most existing energy models are designed to strongly discourage the placement of several vertices on the same position, while clusterings may place many vertices in the same cluster. Technically, existing energy models are not mathematically equivalent to modularity because the exponent of the distance in the repulsion energy is fixed and too small; the (a, r)-energy model is parameterized with this exponent.

Third, the modularity measure and most energy models were originally defined for networks without vertex weights. The vertices are implicitly weighted with 1 in most classic energy models (e.g., [11, 12, 26]), and with their degree in the original modularity measure [8]. It was only recently observed that degree-weighting may also improve the readability and interpretability of energy-based layouts [14, 27]. The definitions of (a, r)-energy and modularity in Sec. II are generalized to arbitrary vertex weights, and thus subsume both degree weights and unit weights.

C. Related work

In the analysis of dissimilarity matrices, the computation of clusterings and layouts with identical quality measures is fairly common (e.g., [28, 29]). The trick is to represent both clusterings and layouts of dissimilarity matrices as dissimilarity matrices: The dissimilarity of two objects in a layout can be defined as their Euclidean distance (as for networks), and the dissimilarity of two objects in a clustering can be defined as the average dissimilarity of the objects in their clusters (unlike for networks, which specify no dissimilarities for their vertices). With this common representation of clusterings and layouts, it is easy to design common quality measures.

For networks, there appear to be no previous proposals of using identical quality measures for both clusterings and layouts. Some clustering algorithms compute layouts as intermediate results, for example eigenvector-based heuristics for modularity clustering [30, 31] and approximation algorithms for some related partitioning problems [32, 33, 34], but these layouts are not intended to be useful on their own.

V. OPTIMAL-ENERGY LAYOUTS CONFORM TO OPTIMAL-MODULARITY CLUSTERINGS

Clusterings and layouts complement each other as representations for the community structure of networks. Layouts are limited to two or three dimensions in practice, and thus cannot faithfully represent inherently high-dimensional structures, but they may show crucial details that are missing in clusterings:

- the density between clusters, and more generally, the relationship between clusters, e.g., whether their separation is clear or fuzzy, and which vertices form their interface,

- the density within clusters, and more generally, the internal structure of clusters, e.g., whether a dense cluster is composed of even denser subclusters,

- the density between vertices and clusters, e.g., whether a vertex is central or peripheral to its cluster, or whether the assignment of a vertex to a cluster is rather arbitrary because it is closely related to several other clusters.

However, a layout only permits these interpretations if it is *consistent* with the respective clustering, i.e., if the layout and the clustering group the vertices according to the same criteria. In previous works, some authors nonetheless consider vertex groups in arbitrary force-directed layouts as clusters, while others rightly note that they have no reasons to suppose that such interpretations are valid. Sections III and IV finally provide such reasons, as summarized in the following subsection.

A. Evidence

Section IV showed that for clusterings with k clusters, considered as restricted $(k-1)$-dimensional layouts, the (a,r)-energy model is equivalent to the modularity measure if $a>-1$ and $r>-1$. Thus (unrestricted) layouts with optimal (a,r)-energy are relaxations of clusterings with optimal modularity if (a) the layouts have at least $k-1$ dimensions, and (b) $a>-1$ and $r>-1$.

Concerning condition (a), the dimensionality of layouts can be somewhat reduced without large changes of the pairwise vertex distances, and thus without large changes of the (a,r)-energy. Hence the consistency of optimal layouts and optimal clusterings does not break down immediately if the layout has less dimensions than the clustering has clusters.

Condition (b) does not imply that layouts with optimal (a,r)-energy closely resemble clusterings with optimal modularity precisely for $a>-1$ and $r>-1$. On the one hand, the condition $r>-1$ is necessary for clusterings to permit the assignment of several vertices to the same cluster, but not for layouts which may group vertices without placing them on exactly the same position. On the other hand, the precise values of a and r hardly matter for clusterings where the distance between vertices is either 0 or 1, but were shown to be important for layouts in Sec. III. Considering the results of Sec. III, (a,r)-energy layouts most closely resemble modularity clusterings if

- $a>r$, $a\geq 0$, and $r\leq 0$ (by the definition of (a,r)-energy),

- $a\approx 0$, such that distances do not reflect path lengths, and

- $a-r\approx 1$, or at least $a-r\not\gg 1$, such that distances reflect densities.

B. Examples

The purpose of this subsection is to illustrate the consistency of (a,r)-energy layouts and modularity clusterings, and the benefits of this consistency, for several real-world networks. It should be stressed that the purpose is *not* to validate the (a,r)-energy model or the modularity measure, which are already widely used and discussed in many previous works; and the purpose is *not* to *prove* the consistency of (a,r)-energy layouts and modularity clusterings, because the mathematical evidence summarized in the previous subsection is more general than any number of examples.

The example networks are listed in Table I. The weight of each vertex is set to its degree, as in the original modularity measure [8] and in the edge-repulsion LinLog energy model [14]. In visualizations, each vertex is represented as a box, its degree (weight) as area of the box, and its cluster membership as shape of the box.

TABLE I: Example networks

Name	Size	Source
Karate Club	34	[35, Figure 3]; unweighted version used in [8, 17, 36]
Book Co-Purchase	105	V. Krebs, provided M. Newman[a]; also used in [17, 36]
Food Classification	45	[37], published in [38, Table 5.1]
World Trade	66	World Bank[b]

[a] `http://www-personal.umich.edu/~mejn/netdata/`
[b] Trade and Production Database at `http://www.worldbank.org`

As motivated in the previous subsection, the parameters of the energy model are set to $a=0$ and $r\in\{-2,-1.5,-1\}$, with $r=-2$ for networks with very nonuniform density (modularity >0.5), and $r=-1$ for networks with fairly uniform density (modularity <0.3). The variation of r improves the readability by ensuring that vertices are not placed too closely, but otherwise does not affect the grouping of the vertices.

Because the exact optimization of (a,r)-energy and modularity is computationally hard, the presented layouts and clusterings are not guaranteed to be optimal (except for the clustering of the Book Co-Purchase network [17]), but are the best known representations. The Java program used for generating these representations is freely available [42]. It employs the Barnes-Hut algorithm for energy minimization, and agglomeration with multi-level refinement for modularity maximization (see Sec. II C).

In the Karate Club network (Fig. 4), each vertex represents a member of a karate club, and the edge weight of each vertex pair specifies the number of contexts (like university classes, bars, or karate tournaments) in which the two members interacted. The main vertex groups in the $(0,-1.5)$-energy layout coincide with the four clusters of the modularity clustering, and the layout correctly indicates that joining triangles and circles into a single

cluster is almost as good as separating them (modularity 0.435 vs. 0.445). The clustering and the layout both segregate the members who left the club after the instructor was fired (gray boxes), with the exception of one member who followed the instructor mainly to preserve his chance for the black belt.

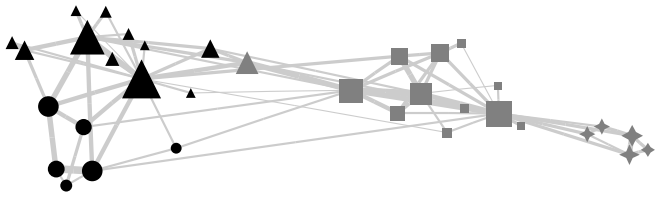

FIG. 4: $(0, -1.5)$-energy layout and modularity clustering (represented by shapes) of the Karate Club network. The modularity of the clustering is 0.445. Gray boxes represent members who left the club after the instructor was fired.

In the Book Co-Purchase network (Fig. 5), the vertices represent books on US politics, and edges of weight 1 connect books that were frequently purchased together. The clusters are generally well-separated in the layout; a few members of the smaller central clusters are placed closely to one of the two large clusters, which correctly indicates that they are densely connected with parts of these large clusters, and their assignment to a smaller cluster is a close decision. The clustering and the layout, especially their two main groups, conform well to Newman's classification [36] of the books as liberal (light gray), neutral (dark gray), or conservative (black); the layout is more suitable to represent the liberal-to-conservative ordering of the books.

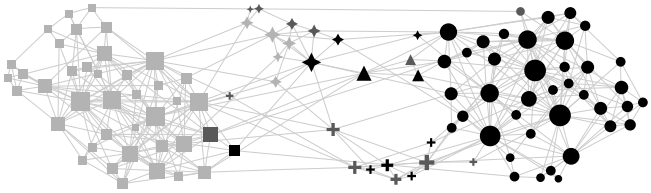

FIG. 5: $(0, -2)$-energy layout and modularity clustering of the Book Co-Purchase network. The modularity is 0.527. Shades represent the classification as liberal (light gray), neutral (dark gray), or conservative (black).

The Food Classification network (Fig. 6) represents the categorizations of 45 foods by 38 subjects of a psychological experiment, who were asked to sort the foods into as many categories as they wished based on perceived similarity. Each vertex represents a food, and the edge weight of each vertex pair is the number of subjects who assigned the corresponding foods to the same category. The clusters correspond well to groups in the layout, but the layout also indicates that the borders between some clusters are rather fuzzy (e.g., between snacks and sweets), that some clusters could be split into subclusters (e.g., fruits and vegetables), and that some foods cannot be clearly assigned to a single cluster (e.g., water, spaghetti). The grouping in both the clustering and the layout largely conforms to common food categories.

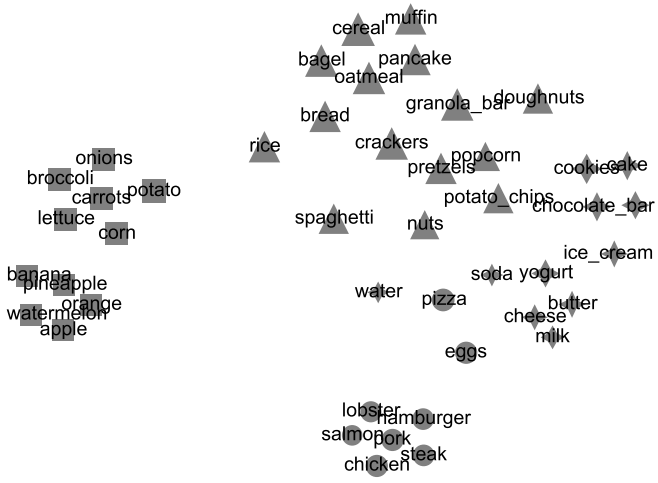

FIG. 6: $(0, -1.5)$-energy layout and modularity clustering of the Food Classification network. The modularity of the clustering is 0.402. (The edges are elided to avoid clutter.)

The World Trade network (Fig. 7) models the trade between 66 countries in the year 1999. The vertices represent countries, and the edge weight of each vertex pair specifies the trade volume between the corresponding countries in US dollar. The clustering and the layout both group the countries of the three major economic areas (East Asia / Australia, America, and Europe). The layout also reflects that countries like IRN and EGY cannot be clearly assigned to either the East Asian or the European group, and shows many smaller groups of closely interlocked countries like CHN and HKG, AUS and NZL, GBR and IRL, and the Nordic countries.

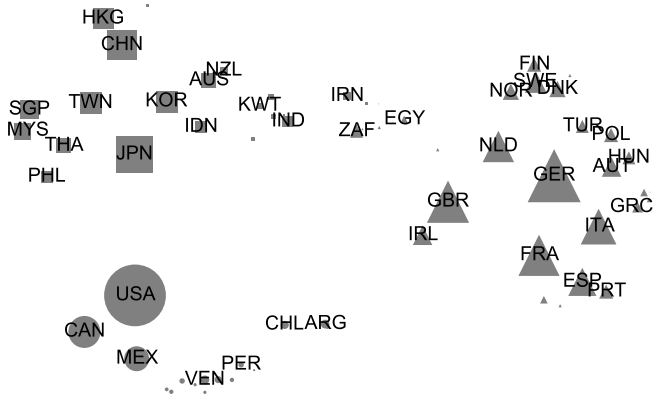

FIG. 7: $(0, -1)$-energy layout and modularity clustering of the World Trade network. The modularity of the clustering is 0.275. (The edges are elided to avoid clutter.)

VI. CONCLUSION

As representations for the community structure of networks, layouts subsume clusterings, thus quality measures for layouts subsume quality measures for clusterings, and in fact prominent existing quality measures for layouts – namely, energy models based on the pairwise attraction and repulsion of vertices – subsume a prominent existing quality measure for clusterings – namely, the modularity measure of Newman and Girvan. This result has implications for the entire lifecycle of quality measures:

- Design: New and existing quality measures for layouts may be applied to clusterings and vice versa. For example, recent extensions of the modularity measure to directed networks [39] and bipartite networks [40] can be directly generalized to energy models for layouts.
- Evaluation: The evaluation of quality measures for clusterings and layouts can be partly unified, i.e., performed without distinguishing between clusterings and layouts. This has been demonstrated in [15] with a computation of the expected measurement value for networks with uniform expected density, a particularly important analysis technique [14, 31, 41].
- Optimization: Components of clustering algorithms may be reused in layout algorithms and vice versa, for example the agglomeration (coarsening) phase of multi-level heuristics. Moreover, energy-based layout algorithms might serve as initial stage of clustering algorithms, similarly to eigenvector-based layout algorithms in existing approaches (see Sec. IV C).
- Application: Unified quality measures help to ensure the consistency of clusterings and layouts (see Sec. V), which is crucial because both representations are often used together.

Acknowledgments

The author thanks Martin Junghans, Randolf Rotta, and Frank Steinbrückner for helpful discussions.

[1] H. A. Simon, Proc. Am. Philos. Soc. **106**, 467 (1962).
[2] U. Brandes, in *Drawing Graphs: Methods and Models*, edited by M. Kaufmann and D. Wagner (Springer, Berlin, 2001), LNCS 2025, pp. 71–86.
[3] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs* (Prentice Hall, Upper Saddle River, 1999).
[4] M. E. J. Newman, Eur. Phys. J. B **38**, 321 (2004).
[5] M. Gaertler, in *Network Analysis: Methodological Foundations*, edited by U. Brandes and T. Erlebach (Springer, Berlin, 2005), LNCS 3418, pp. 178–215.
[6] S. E. Schaeffer, Comput. Sci. Rev. **1**, 27 (2007).
[7] S. Fortunato and C. Castellano (2007), preprint arXiv:0712.2716.
[8] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[9] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer, New York, 1997).
[10] T. Kamada and S. Kawai, Inform. Process. Lett. **31**, 7 (1989).
[11] T. M. J. Fruchterman and E. M. Reingold, Software Pract. Exper. **21**, 1129 (1991).
[12] R. Davidson and D. Harel, ACM Trans. Graphics **15**, 301 (1996).
[13] A. Noack, in *Proceedings of the 11th International Symposium on Graph Drawing (GD 2003)*, edited by G. Liotta (Springer, Berlin, 2004), LNCS 2912, pp. 425–436.
[14] A. Noack, J. Graph Algorithms Appl. **11**, 453 (2007).
[15] A. Noack, Ph.D. thesis, Brandenburg University of Technology (2007), URL `http://nbn-resolving.de/urn:nbn:de:kobv:co1-opus-40663`.
[16] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).
[17] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE Trans. Knowl. Data Eng. **20**, 172 (2008).
[18] S. Hachul and M. Jünger, J. Graph Algorithms Appl. **11**, 345 (2007).
[19] J. Barnes and P. Hut, Nature **324**, 446 (1986).
[20] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech., P09008 (2005).
[21] P. Schuetz and A. Caflisch, Phys. Rev. E **77**, 046112 (2008).
[22] R. Rotta, Diploma thesis, Brandenburg University of Technology (2008), URL `http://www.informatik.tu-cottbus.de/~rrotta/`.
[23] R. Diestel, *Graph Theory* (Springer, New York, 2000), 2nd ed.
[24] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).
[25] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[26] P. Eades, Congressus Numerantium **42**, 149 (1984).
[27] A. Noack, in *Proceedings of the 13th International Symposium on Graph Drawing (GD 2005)*, edited by P. Healy and N. S. Nikolov (Springer, Berlin, 2006), LNCS 3843, pp. 309–320.
[28] R. N. Shepard, Science **210**, 390 (1980).
[29] J. D. Carroll and S. Pruzansky, in *Similarity and Choice: Papers in Honour of Clyde Coombs*, edited by E. D. Lantermann and H. Feger (Huber, Bern, 1980), pp. 108–139.
[30] S. White and P. Smyth, in *Proceedings of the 5th SIAM International Conference on Data Mining (SDM 2005)* (SIAM, Philadelphia, 2005), pp. 274–285.
[31] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[32] N. Linial, E. London, and Y. Rabinovich, Combinatorica **15**, 215 (1995).

[33] Y. Aumann and Y. Rabani, SIAM J. Comput. **27**, 291 (1998).
[34] S. Arora, S. Rao, and U. V. Vazirani, in *Proceedings of the 36th ACM Symposium on Theory of Computing (STOC 2004)*, edited by L. Babai (ACM, New York, 2004), pp. 222–231.
[35] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[36] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[37] B. H. Ross and G. L. Murphy, Cognitive Psychol. **38**, 495 (1999).
[38] L. Hubert, P. Arabie, and J. Meulman, *Combinatorial Data Analysis: Optimization by Dynamic Programming* (SIAM, Philadelphia, 2001).
[39] E. A. Leicht and M. E. J. Newman, Phys. Rev. Lett. **100**, 118703 (2008).
[40] M. J. Barber, Phys. Rev. E **76**, 066102 (2007).
[41] M. Gaertler, R. Görke, and D. Wagner, in *Proceedings of the 3rd International Conference on Algorithmic Aspects in Information and Management (AAIM 2007)*, edited by M.-Y. Kao and X.-Y. Li (Springer, Berlin, 2007), LNCS 4508, pp. 11–26.
[42] `http://code.google.com/p/linloglayout/`